upsetplot Documentation

Release 0.3.0.post3

Joel Nothman

Jul 04, 2019

Contents

This is another Python implementation of UpSet plots by Lex et al. [Lex2014]. UpSet plots are used to visualise set overlaps; like Venn diagrams but more readable. Documentation is at https://upsetplot.readthedocs.io.

This `upsetplot` library tries to provide a simple interface backed by an extensible, object-oriented design.

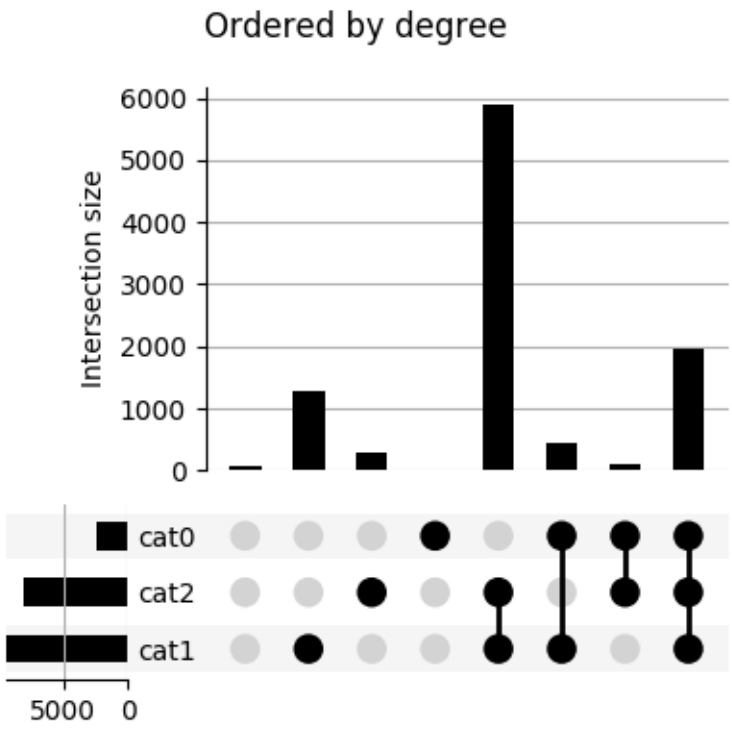
The basic input format is a `pandas.Series` containing counts corresponding to subset sizes, where each subset is an intersection of named categories. The index of the Series indicates which rows pertain to which categories, by having multiple boolean indices, like `example` in the following:

```
>>> from upsetplot import generate_counts
>>> example = generate_counts()
>>> example
cat0   cat1   cat2
False  False  False      56
              True      283
       True   False    1279
              True     5882
True   False  False      24
              True       90
       True   False     429
              True     1957
Name: value, dtype: int64
```

Then:

```
>>> from upsetplot import plot
>>> plot(example)
>>> from matplotlib import pyplot
>>> pyplot.show()
```

makes:

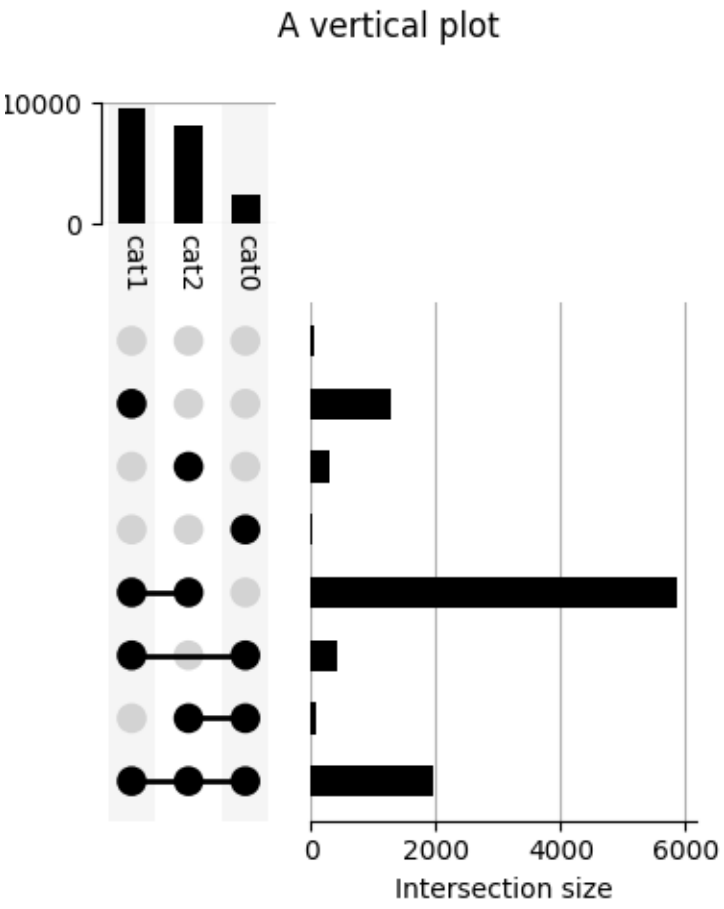


This plot shows the cardinality of every category combination seen in our data. The leftmost column counts items absent from any category. The next three columns count items only in `cat1`, `cat2` and `cat3` respectively, with following columns showing cardinalities for items in each combination of exactly two named sets. The rightmost column counts items in all three sets.

CHAPTER 1

Rotation

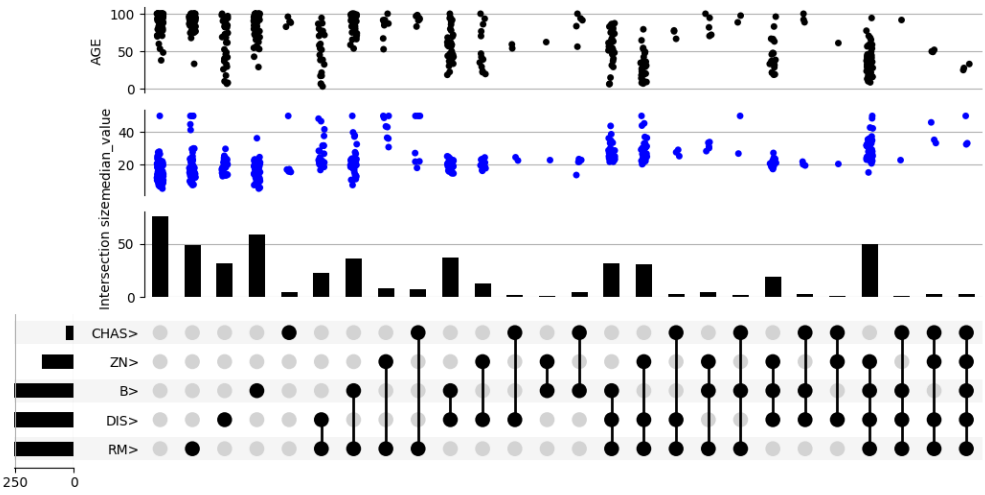
We call the above plot style “horizontal” because the category intersections are presented from left to right. Vertical plots are also supported!



CHAPTER 2

Distributions

Providing a DataFrame rather than a Series as input allows us to expressively plot the distribution of variables in each subset.



CHAPTER 3

Loading datasets

While the dataset above is randomly generated, you can prepare your own dataset for input to upsetplot. A helpful tool is *from_memberships*, which allows us to reconstruct the example above by indicating each data point's category membership:

```
>>> from upsetplot import from_memberships
>>> example = from_memberships(
...     [[],
...      ['cat2'],
...      ['cat1'],
...      ['cat1', 'cat2'],
...      ['cat0'],
...      ['cat0', 'cat2'],
...      ['cat0', 'cat1'],
...      ['cat0', 'cat1', 'cat2'],
...      ],
...      data=[56, 283, 1279, 5882, 24, 90, 429, 1957]
... )
>>> example
cat0   cat1   cat2
False  False  False      56
               True      283
       True   False    1279
               True     5882
True   False  False      24
               True       90
       True   False     429
               True     1957
dtype: int64
```

See also *from_contents*, another way to describe categorised data.

3.1 Installation

To install the library, you can use `pip`:

```
$ pip install upsetplot
```

Installation requires:

- pandas
- matplotlib >= 2.0
- seaborn to use *`UpSet.add_catplot`*

It should then be possible to:

```
>>> import upsetplot
```

in Python.

3.2 Why an alternative to py-upset?

Probably for petty reasons. It appeared py-upset was not being maintained. Its input format was undocumented, inefficient and, IMO, inappropriate. It did not facilitate showing plots of each subset's distribution as in Lex et al's work introducing UpSet plots. Nor did it include the horizontal bar plots illustrated there. It did not support Python 2. I decided it would be easier to construct a cleaner version than to fix it.

3.3 References

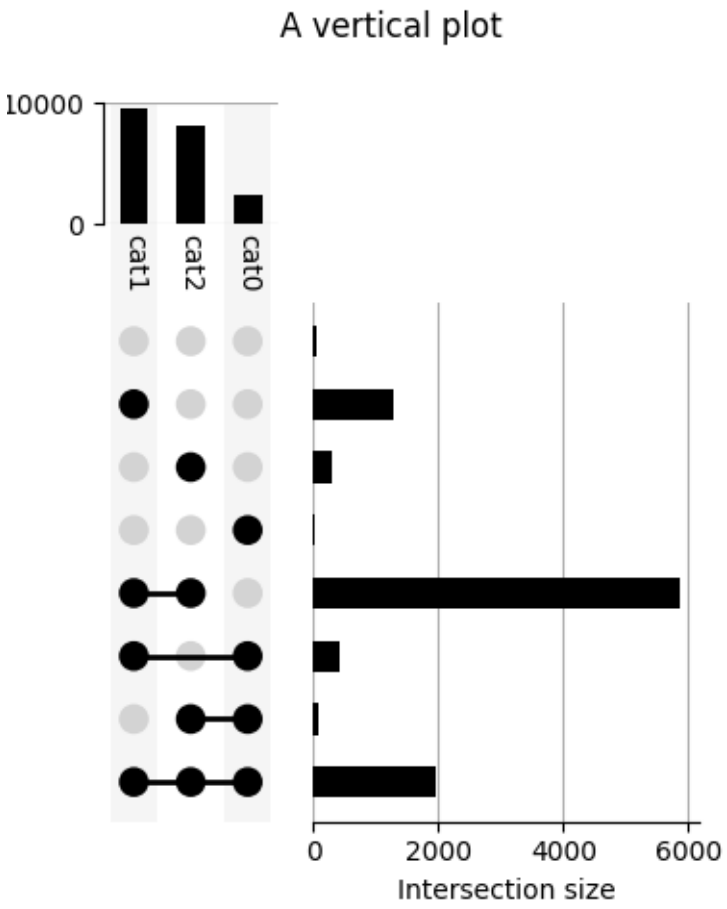
3.3.1 Examples

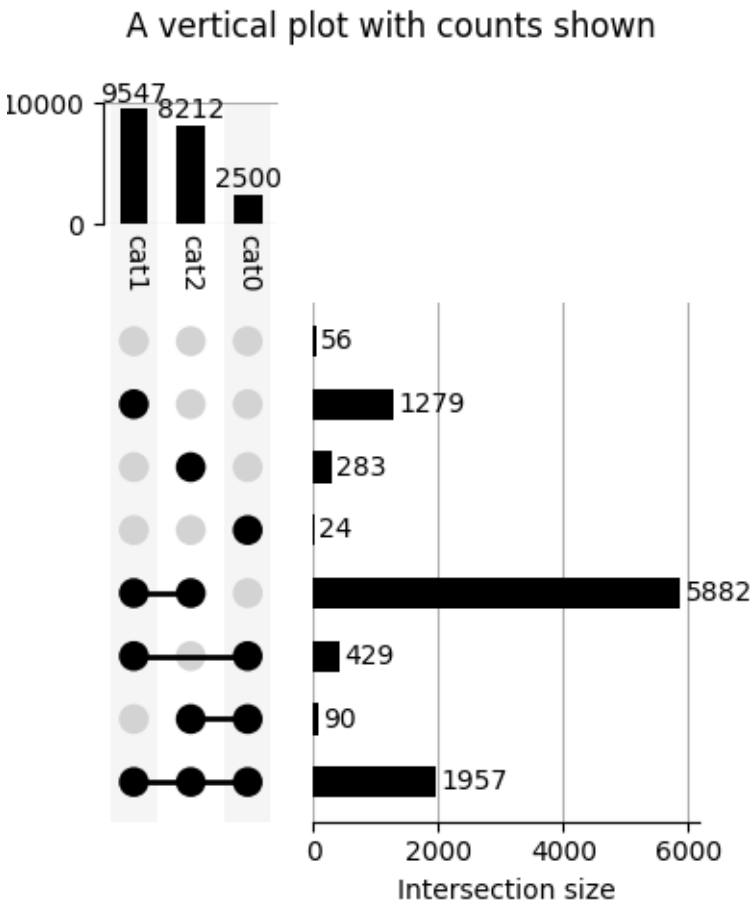
Introductory examples for upsetplot.

Note: Click *here* to download the full example code

Vertical orientation

This illustrates the effect of orientation='vertical'.


A vertical plot
10000
0
cat1
cat2
cat0
0
2000
4000
6000
Intersection size



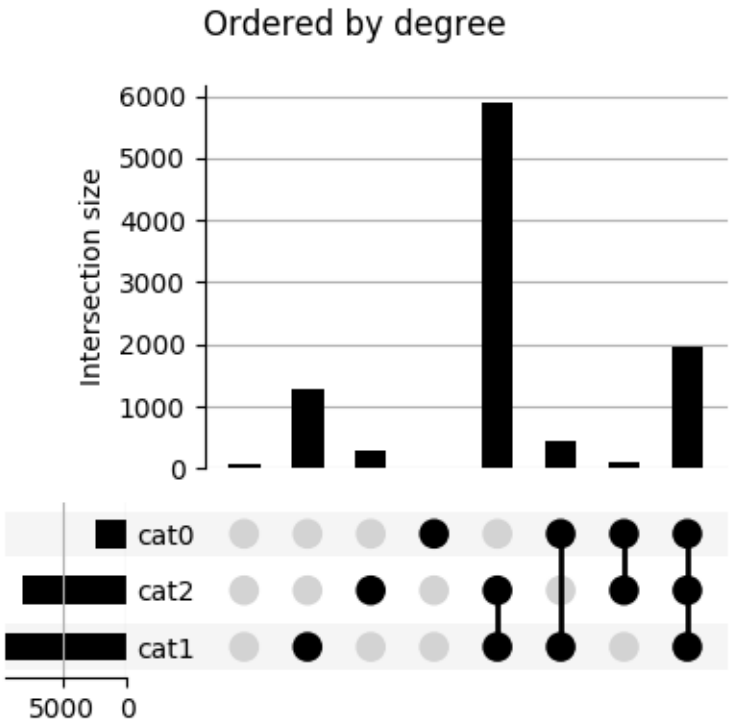

```
from matplotlib import pyplot as plt
from upsetplot import generate_counts, plot

example = generate_counts()
plot(example, orientation='vertical')
plt.suptitle('A vertical plot')
plt.show()

plot(example, orientation='vertical', show_counts='%d')
plt.suptitle('A vertical plot with counts shown')
plt.show()
```

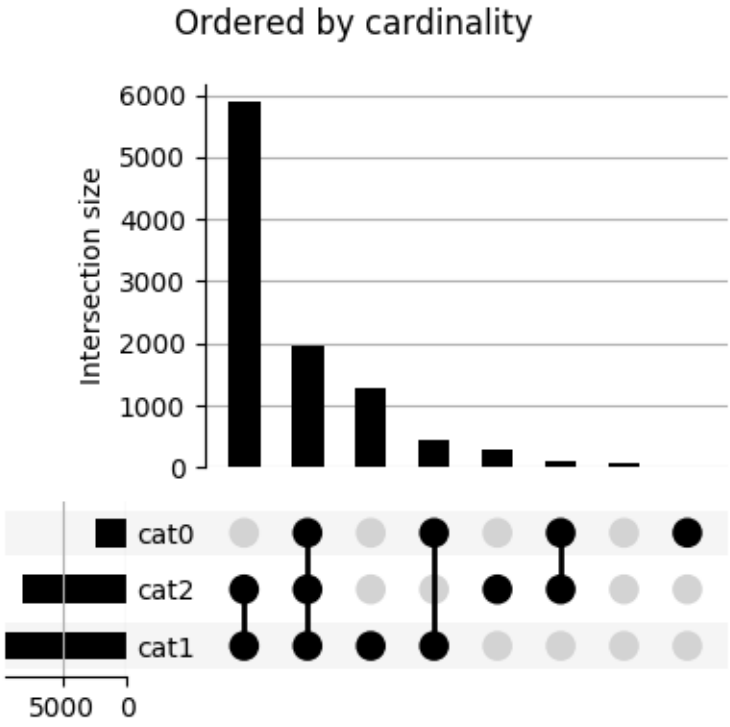
Total running time of the script: (0 minutes 0.550 seconds)

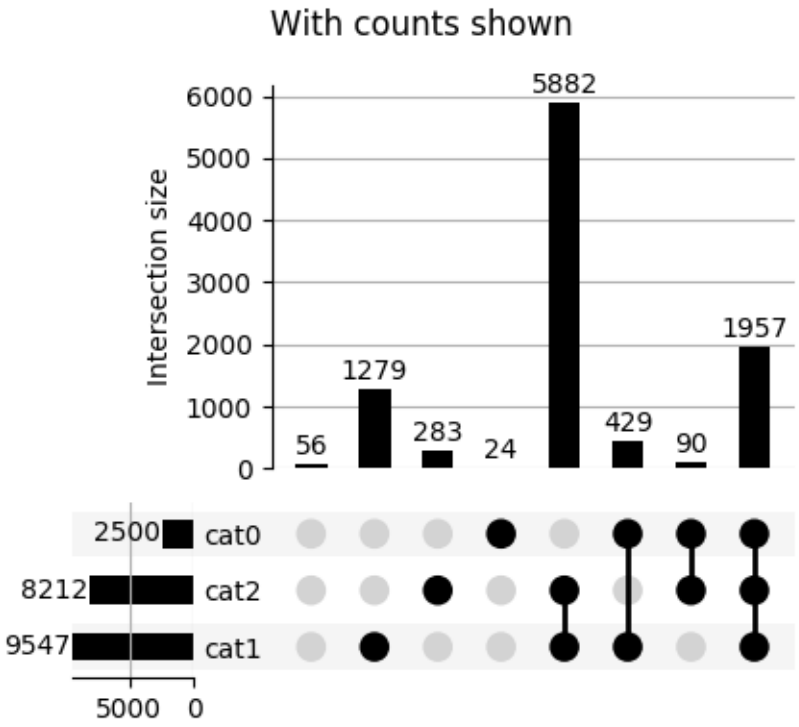
Note: Click *here* to download the full example code

Plotting with generated data

This example illustrates basic plotting functionality using generated data.


Ordered by degree
Intersection size
6000
5000
4000
3000
2000
1000
0
cat0
cat2
cat1
5000
0


Ordered by cardinality
Intersection size
6000
5000
4000
3000
2000
1000
0
cat0
cat2
cat1
5000
0



-

Out:

```
cat0   cat1   cat2
False  False  False      56
               True      283
       True   False    1279
               True     5882
True   False  False      24
               True       90
       True   False     429
               True     1957
Name: value, dtype: int64
```

```
from matplotlib import pyplot as plt
from upsetplot import generate_counts, plot

example = generate_counts()
print(example)

plot(example)
plt.suptitle('Ordered by degree')
plt.show()

plot(example, sort_by='cardinality')
plt.suptitle('Ordered by cardinality')
plt.show()
```

(continues on next page)

(continued from previous page)

```
plot(example, show_counts='%d')
plt.suptitle('With counts shown')
plt.show()
```

Total running time of the script: (0 minutes 0.748 seconds)

Note: Click *here* to download the full example code

Above-average features in Boston

Explore above-average neighborhood characteristics in the Boston dataset.

Here we take some features correlated with house price, and look at the distribution of median house price when each of these features is above average.

The most correlated features are:

ZN proportion of residential land zoned for lots over 25,000 sq.ft.

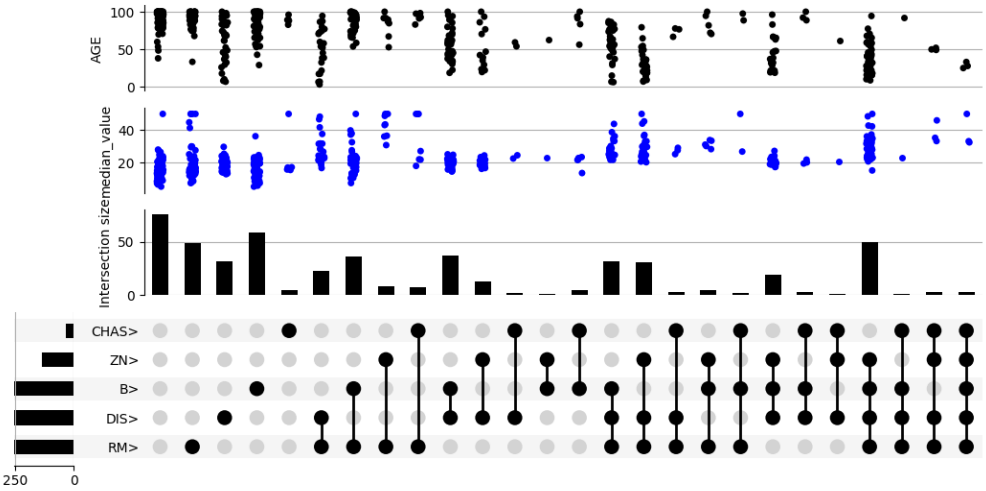
CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

RM average number of rooms per dwelling

DIS weighted distances to five Boston employment centres

B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

This kind of dataset analysis may not be a practical use of UpSet, but helps to illustrate the `UpSet.add_catplot()` feature.




```
import pandas as pd
from sklearn.datasets import load_boston
from matplotlib import pyplot as plt
from upsetplot import UpSet
```

(continues on next page)

(continued from previous page)

```
# Load the dataset into a DataFrame
boston = load_boston()
boston_df = pd.DataFrame(boston.data, columns=boston.feature_names)

# Get five features most correlated with median house value
correls = boston_df.corrwith(pd.Series(boston.target),
                             method='spearman').sort_values()
top_features = correls.index[-5:]

# Get a binary indicator of whether each top feature is above average
boston_above_avg = boston_df > boston_df.median(axis=0)
boston_above_avg = boston_above_avg[top_features]
boston_above_avg = boston_above_avg.rename(columns=lambda x: x + '>')

# Make this indicator mask an index of boston_df
boston_df = pd.concat([boston_df, boston_above_avg],
                      axis=1)
boston_df = boston_df.set_index(list(boston_above_avg.columns))

# Also give us access to the target (median house value)
boston_df = boston_df.assign(median_value=boston.target)

# UpSet plot it!
upset = UpSet(boston_df, subset_size='count', intersection_plot_elements=3)
upset.add_catplot(value='median_value', kind='strip', color='blue')
upset.add_catplot(value='AGE', kind='strip', color='black')
upset.plot()
plt.show()
```

Total running time of the script: (0 minutes 1.095 seconds)

3.3.2 API Reference

Plotting

upsetplot.**plot**(*data*, *fig=None*, ***kwargs*)

Make an UpSet plot of data on fig

Parameters

data [pandas.Series or pandas.DataFrame] Values for each set to plot. Should have multi-index where each level is binary, corresponding to set membership. If a DataFrame, `sum_over` must be a string or False.

fig [matplotlib.figure.Figure, optional] Defaults to a new figure.

kwargs Other arguments for *UpSet*

Returns

subplots [dict of matplotlib.axes.Axes] Keys are 'matrix', 'intersections', 'totals', 'shading'

class upsetplot.**UpSet**(*data*, *orientation='horizontal'*, *sort_by='degree'*, *sort_categories_by='cardinality'*, *subset_size='legacy'*, *sum_over=None*, *facecolor='black'*, *with_lines=True*, *element_size=32*, *intersection_plot_elements=6*, *totals_plot_elements=2*, *show_counts=''*, *sort_sets_by='deprecated'*)

Manage the data and drawing for a basic UpSet plot

Primary public method is *plot()*.

Parameters

data [pandas.Series or pandas.DataFrame] Elements associated with categories (a DataFrame), or the size of each subset of categories (a Series). Should have MultiIndex where each level is binary, corresponding to category membership. If a DataFrame, `sum_over` must be a string or False.

orientation [{'horizontal' (default), 'vertical'}] If horizontal, intersections are listed from left to right.

sort_by [{'cardinality', 'degree'}] If 'cardinality', subset are listed from largest to smallest. If 'degree', they are listed in order of the number of categories intersected.

sort_categories_by [{'cardinality', None}] Whether to sort the categories by total cardinality, or leave them in the provided order.

subset_size [{'auto', 'count', 'sum'}] Configures how to calculate the size of a subset. Choices are:

'auto' If `data` is a DataFrame, count the number of rows in each group, unless `sum_over` is specified. If `data` is a Series with at most one row for each group, use the value of the Series. If `data` is a Series with more than one row per group, raise a ValueError.

'count' Count the number of rows in each group.

'sum' Sum the value of the `data` Series, or the DataFrame field specified by `sum_over`.

Until version 0.4, the default is 'legacy' which uses `sum_over` to control this behaviour. From version 0.4, 'auto' will be default.

sum_over [str or None] If `subset_size='sum'` or `'auto'`, then the intersection size is the sum of the specified field in the `data` DataFrame. If a Series, only None is supported and its value is summed.

If `subset_size='legacy'`, `sum_over` must be specified when `data` is a DataFrame. If False, the intersection plot will show the count of each subset. Otherwise, it shows the sum of the specified field.

facecolor [str] Color for bar charts and dots.

with_lines [bool] Whether to show lines joining dots in the matrix, to mark multiple categories being intersected.

element_size [float or None] Side length in pt. If None, size is estimated to fit figure

intersection_plot_elements [int] The intersections plot should be large enough to fit this many matrix elements.

totals_plot_elements [int] The totals plot should be large enough to fit this many matrix elements.

show_counts [bool or str, default=False] Whether to label the intersection size bars with the cardinality of the intersection. When a string, this formats the number. For example, '%d' is equivalent to True.

sort_sets_by

Methods

add_catplot(self, kind[, value, elements])	Add a seaborn catplot over subsets when *plot()* is called.
make_grid(self[, fig])	Get a SubplotSpec for each Axes, accounting for label text width
plot(self[, fig])	Draw all parts of the plot onto fig or a new figure
plot_intersections(self, ax)	Plot bars indicating intersection size
plot_matrix(self, ax)	Plot the matrix of intersection indicators onto ax
plot_totals(self, ax)	Plot bars indicating total set size

plot_shading

add_catplot (*self*, *kind*, *value=None*, *elements=3*, ***kw*)
Add a seaborn catplot over subsets when *plot()* is called.

Parameters

kind [str] One of {"point", "bar", "strip", "swarm", "box", "violin", "boxen"}

value [str, optional] Column name for the value to plot (i.e. y if orientation='horizontal'), required if `data` is a DataFrame.

elements [int, default=3] Size of the axes counted in number of matrix elements.

****kw** [dict] Additional keywords to pass to `seaborn.catplot()`.

Our implementation automatically determines 'ax', 'data', 'x', 'y' and 'orient', so these are prohibited keys in `kw`.

Returns

None

make_grid (*self*, *fig=None*)
Get a SubplotSpec for each Axes, accounting for label text width

plot (*self*, *fig=None*)
Draw all parts of the plot onto fig or a new figure

Parameters

fig [matplotlib.figure.Figure, optional] Defaults to a new figure.

Returns

subplots [dict of matplotlib.axes.Axes] Keys are 'matrix', 'intersections', 'totals', 'shading'

plot_intersections (*self*, *ax*)
Plot bars indicating intersection size

plot_matrix (*self*, *ax*)
Plot the matrix of intersection indicators onto ax

plot_totals (*self*, *ax*)
Plot bars indicating total set size

Dataset loading and generation

upsetplot.**from_contents**(*contents*, *data=None*, *id_column='id'*)
Build data from category listings

Parameters

contents [Mapping (or iterable over pairs) of strings to sets] Keys are category names, values are sets of identifiers (int or string).

data [DataFrame, optional] If provided, this should be indexed by the identifiers used in `Python Documentation contents`.

id_column [str, default='id'] The column name to use for the identifiers in the output.

Returns

DataFrame `data` is returned with its index indicating category membership, including a column named according to id_column. If data is not given, the order of rows is not assured.

Notes

The order of categories in the output DataFrame is determined from `Python Documentation contents`, which may have non-deterministic iteration order.

Examples

```
>>> from upsetplot import from_contents
>>> contents = {'cat1': ['a', 'b', 'c'],
...             'cat2': ['b', 'd'],
...             'cat3': ['e']}
>>> from_contents(contents)  # doctest: +NORMALIZE_WHITESPACE
                  id
cat1  cat2  cat3
True  False False  a
      True  False  b
      False False  c
False True  False  d
      False True   e
>>> import pandas as pd
>>> contents = {'cat1': [0, 1, 2],
...             'cat2': [1, 3],
...             'cat3': [4]}
>>> data = pd.DataFrame({'favourite': ['green', 'red', 'red',
...                                    'yellow', 'blue']})
>>> from_contents(contents, data=data)  # doctest: +NORMALIZE_WHITESPACE
                   id favourite
cat1  cat2  cat3
True  False False   0     green
      True  False   1       red
      False False   2       red
False True  False   3    yellow
      False True    4      blue
```

upsetplot.**from_memberships**(*memberships*, *data=None*)
Load data where each sample has a collection of category names

The output should be suitable for passing to *`UpSet`* or *`plot`*.

Parameters

memberships [sequence of collections of strings] Each element corresponds to a data point, indicating the sets it is a member of. Each category is named by a string.

data [Series-like or DataFrame-like, optional] If given, the index of category memberships is attached to this data. It must have the same length as `memberships`. If not given, the series will contain the value 1.

Returns

DataFrame or Series `data` is returned with its index indicating category membership. It will be a Series if `data` is a Series or 1d numeric array. The index will have levels ordered by category names.

Examples

```
>>> from upsetplot import from_memberships
>>> from_memberships([
...     ['cat1', 'cat3'],
...     ['cat2', 'cat3'],
...     ['cat1'],
...     []
... ])  # doctest: +ELLIPSIS, +NORMALIZE_WHITESPACE
cat1   cat2   cat3
True   False  True     1
False  True   True     1
True   False  False    1
False  False  False    1
Name: ones, dtype: ...
>>> # now with data:
>>> import numpy as np
>>> from_memberships([
...     ['cat1', 'cat3'],
...     ['cat2', 'cat3'],
...     ['cat1'],
...     []
... ], data=np.arange(12).reshape(4, 3))  # doctest: +NORMALIZE_WHITESPACE
                   0   1   2
cat1  cat2  cat3
True  False True   0   1   2
False True  True   3   4   5
True  False False  6   7   8
False False False  9  10  11
```

`upsetplot.`**`generate_counts`**(*seed=0*, *n_samples=10000*, *n_categories=3*)

Generate artificial counts corresponding to set intersections

Parameters

seed [int] A seed for randomisation

n_samples [int] Number of samples to generate statistics over

n_categories [int] Number of categories (named "cat0", "cat1", ...) to generate

Returns

Series Counts indexed by boolean indicator mask for each category.

See also:

generate_samples Generates a DataFrame of samples that these counts are derived from.

upsetplot.**generate_samples**(*seed=0*, *n_samples=10000*, *n_categories=3*)
Generate artificial samples assigned to set intersections

Parameters

seed [int] A seed for randomisation

n_samples [int] Number of samples to generate

n_categories [int] Number of categories (named "cat0", "cat1", ...) to generate

Returns

DataFrame Field 'value' is a weight or score for each element. Field 'index' is a unique id for each element. Index includes a boolean indicator mask for each category.

Note: Further fields may be added in future versions.

See also:

generate_counts Generates the counts for each subset of categories corresponding to these samples.

3.3.3 Changelog

What's new in version 0.3

- Added *from_contents* to provide an alternative, intuitive way of specifying category membership of elements.
- To improve code legibility and intuitiveness, `sum_over=False` was deprecated and a `subset_size` parameter was added. It will have better default handling of DataFrames after a short deprecation period.
- `generate_data` has been replaced with *generate_counts* and *generate_samples*.
- Fixed the display of the "intersection size" label on plots, which had been missing.
- Trying to improve nomenclature, upsetplot now avoids "set" to refer to the top-level sets, which are now to be known as "categories". This matches the intuition that categories are named, logical groupings, as opposed to "subsets". To this end:
  - *generate_counts* (formerly `generate_data`) now names its categories "cat1", "cat2" etc. rather than "set1", "set2", etc.
  - the `sort_sets_by` parameter has been renamed to `sort_categories_by` and will be removed in version 0.4.

What's new in version 0.2.1

- Return a Series (not a DataFrame) from *from_memberships* if data is 1-dimensional.

What's new in version 0.2

- Added *from_memberships* to allow a more convenient data input format.
- *plot* and *UpSet* now accept a `pandas.DataFrame` as input, if the `sum_over` parameter is also given.
- Added an *add_catplot* method to *UpSet* which adds Seaborn plots of set intersection data to show more than just set size or total.

- Shading of subset matrix is continued through to totals.
- Added a `show_counts` option to show counts at the ends of bar plots. (#5)
- Defined `_repr_html_` so that an *UpSet* object will render in Jupyter notebooks. (#36)
- Fix a bug where an error was raised if an input set was empty.

Bibliography

[Lex2014] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, Hanspeter Pfister, *UpSet: Visualization of Intersecting Sets*, IEEE Transactions on Visualization and Computer Graphics (InfoVis '14), vol. 20, no. 12, pp. 1983–1992, 2014. doi: doi.org/10.1109/TVCG.2014.2346248

Index